

# Calibrating Survey Weights in Stata

Jeff Pitblado

StataCorp LLC

2018 Canadian Stata Users Group Meeting  
Vancouver, Canada

# Outline

Motivation

Methods

Syntax

Stata Example

Summary

# Motivation

## Survey data analysis

We collect data from a population of interest so that we can describe the population and make inferences about the population.

## Sampling

The goal of sampling is to collect data that represents the population of interest.

- ▶ If the sample does not reasonably represent the population of interest, then we cannot accurately describe the population or make inferences.

# Motivation

## Weighting

Sampling weights provide a measure of how many individuals a given sampled observation represents in the population.

- ▶ In simple random sampling (SRS), the sampling weight is constant

$$w_i = N/n$$

- ▶  $N$  is the population size
- ▶  $n$  is the sample size
- ▶ Other, more complicated, sampling designs are also self weighting, but this is more a special case than the norm.

# Motivation

## Weighting

Survey methods employ sampling weights, in the computation of descriptive statistics and the fitting of regression models, in order to describe the population and make inferences about the population.

## Sampling weights

- ▶ Correctly scaled sampling weights are necessary for estimating population totals.
- ▶ Typically provide for consistent and approximately unbiased estimates.
- ▶ Typically provide for more accurate variance estimation when used with the survey design characteristics.

# Motivation

## Non-response

Failure to observe all the individuals that were selected for the sample.

- ▶ A common cause for some groups to be under-represented and other groups to be over-represented.

# Motivation

## Example

Consider a survey design that intends for individuals sampled from group  $g$  to have weight

$$w_{gi} = \frac{N_g}{n_g}$$

- ▶  $N_g$  is the population size for group  $g$
- ▶  $n_g$  is the group's sample size

If we observe  $m_g < n_g$  individuals, then  $w_{gi}$  is smaller than it should be. Group  $g$  is under-represented in the sample.

- ▶ Seems reasonable to adjust  $w_{gi}$  by something that will make them sum to  $N_g$  in the sample.

$$\tilde{w}_{gi} = w_{gi} \frac{n_g}{m_g} = \frac{N_g}{m_g}$$

# Motivation

## Weight adjustment

Weight adjustment tries to give more weight to under-represented groups and less weight to over-represented groups.

- ▶ The idea is to cut down on bias, thus make point estimates more consistent for the things they are estimating.
- ▶ Has been used to force estimation results to be numerically consistent with externally sourced measurements.
- ▶ Tends to result in more efficient point estimates.
  - ▶ The degree to which they are more efficient is a function of the correlation between the analysis variable and the auxiliary information used to adjust the weights.



# Methods

## Poststratification

Adjust weights so that the poststratum totals agree with “known” values.

- ▶ simple method for weight adjustment
- ▶ requires poststratum identifiers are present in the sample information
  - ▶ single categorical auxiliary variable
- ▶ requires population poststratum totals
- ▶ adjustment is a function of the sampling weights and poststratum totals
- ▶ new feature in Stata 9

# Methods

## Calibration

Adjust the sampling weights to minimize the difference between “known” population totals and their weighted estimates.

- ▶ poststratification is a special case
- ▶ supports multiple categorical auxiliary variables
- ▶ supports count and continuous auxiliary variables
- ▶ adjustment is a function of the sampling weights and auxiliary information
- ▶ new feature in Stata 15
  - ▶ raking-ratio method
  - ▶ general regression method (GREG)

# Syntax

## Familiar work flow

1. Use **svyset** to specify the survey design characteristics.
  - ▶ Sampling units
  - ▶ Sampling and replication weights
  - ▶ Strata
  - ▶ Finite population correction (FPC)
  - ▶ Poststratification, raking-ratio, or GREG
2. Use the **svy**: prefix for estimation.
  - ▶ Calibration is supported by the following variance estimation methods:
    - ▶ Linearization
    - ▶ Balanced repeated replication (BRR)
    - ▶ Bootstrap
    - ▶ Jackknife
    - ▶ Successive difference replication (SDR)

# Syntax

```
svyset psu [weight], options || ...
```

## Poststratification options

- ▶ **poststrata** (*varname*) specifies variable containing the poststratum identifiers
- ▶ **postweight** (*varname*) specifies variable containing the poststratum totals

# Syntax

```
svyset psu [weight], options || ...
```

## Calibration options

- ▶ **rake**(*calspec*) specifies the raking-ratio method
- ▶ **regress**(*calspec*) specifies the GREG method
- ▶ *calspec* has syntax

```
varlist, totals(totals)
```

- ▶ *varlist* contains the list of auxiliary variables and allows factor variables notation
- ▶ *totals* specifies the population totals for each auxiliary variable
  - ▶ *var=#* specify each population total separately
  - ▶ *matname* specify the population totals using a matrix

# Stata Example

## Simulated population

frame	count	index	variable
strata	2	<i>h</i>	<b>st1</b>
PSU	1,000	<i>i</i>	<b>su1</b>
SSU	100	<i>j</i>	
total	200,000		

- ▶ **y** is the measurement of interest
- ▶  $\mu_y$ , the mean of **y**, is the parameter of interest
- ▶ **a** and **b** are continuous auxiliary variables
- ▶ **f** and **g** are categorical auxiliary variables

# Stata Example

## Simulated population

$$\mathbf{a}_{hij} = \mu_a + \nu_{a_{hi}} + \epsilon_{a_{hij}}$$

- ▶  $\nu_{a_{hi}}$  i.i.d.  $N(0, 100)$
- ▶  $\epsilon_{a_{hij}}$  i.i.d.  $N(0, 100)$
- ▶  $\nu$  and  $\epsilon$  are independent
- ▶  $\mathbf{a}$  has intraclass correlation  $\rho_a^2 = .5$
- ▶  $\mu_a = 10$
- ▶ total for  $\mathbf{a}$  is 2,000,000
- ▶  $\mathbf{f}$  categorizes  $\mathbf{a}$  into 4 roughly-equal groups

# Stata Example

## Simulated population

$$\mathbf{b}_{hij} = \mu_b + \nu_{b_{hi}} + \epsilon_{b_{hij}}$$

- ▶  $\nu_{b_{hi}}$  i.i.d.  $N(0, 100)$
- ▶  $\epsilon_{b_{hij}}$  i.i.d.  $N(0, 300)$
- ▶  $\nu$  and  $\epsilon$  are independent
- ▶  $\mathbf{b}$  has intraclass correlation  $\rho_b^2 = .25$
- ▶  $\mu_b = 5$
- ▶ total for  $\mathbf{b}$  is 1,000,000
- ▶  $\mathbf{g}$  categorizes  $\mathbf{b}$  into 2 roughly-equal groups



# Stata Example

## Simulated population

Cell and margin sizes of **f** and **g**:

```
. table f g, row col
```

f	g		Total
	1	2	
1	23,238	22,693	45,931
2	25,286	29,486	54,772
3	27,618	25,059	52,677
4	22,615	24,005	46,620
Total	98,757	101,243	200,000

# Stata Example

## Simulated population

$$\mathbf{y}_{hij} = \beta_0 + \beta_1 \mathbf{a}_{hij} + \beta_2 \mathbf{b}_{hij} + \nu_{y_{hi}} + \epsilon_{y_{hij}}$$

- ▶  $\nu_{y_{hi}}$  i.i.d.  $N(0, 100)$
- ▶  $\epsilon_{y_{hij}}$  i.i.d.  $N(0, 100)$
- ▶  $\nu$  and  $\epsilon$  are independent
- ▶  $\mathbf{y}$  has intraclass correlation  $\rho_b^2 = .5$
- ▶  $\beta_0 = 10, \beta_1 = 4, \beta_2 = 2$
- ▶  $\mathbf{y}$  has overall mean

$$\begin{aligned}\mu_y &= \beta_0 + \beta_1 \mu_a + \beta_2 \mu_b \\ &= 10 + 4 \times 10 + 2 \times 5 = 60\end{aligned}$$

# Stata Example

## Simulated population

Strength of association between **y**, **a**, and **b**:

```
. correlate y a b  
(obs=200,000)
```

	y	a	b
y	1.0000		
a	0.8012	1.0000	
b	0.5655	0.0017	1.0000

# Stata Example

## Simulated population

Strength of association between **y**, **f**, and **g**:

```
. correlate y f g  
(obs=200,000)
```

	y	f	g
y	1.0000		
f	0.5774	1.0000	
g	0.2560	-0.0022	1.0000

# Stata Example

## Sample from the population

Stratified two-stage design:

1. select 20 PSUs within each stratum
2. select 10 individuals within each sampled PSU

With zero non-response, this sampling scheme yielded:

- ▶ 400 sampled individuals
- ▶ constant sampling weights

$$pw = 500$$

Other variables:

- ▶ **w4f** – poststratum weights for **f**
- ▶ **w4g** – poststratum weights for **g**

# Stata Example

## Sample weighted cell totals for f

```
. table f [pw=pw], c(freq min w4f) format(%9.0gc)
```

f	Freq.	min(w4f)
1	50,000	45,931
2	75,000	54,772
3	59,000	52,677
4	16,000	46,620

- ▶ Over-represented: 2
- ▶ Under-represented: 4

# Stata Example

## Sample weighted cell totals for **g**

```
. table g [pw=pw], c(freq min w4g) format(%9.0gc)
```

g	Freq.	min(w4g)
1	105,000	98,757
2	95,000	101,243

# Stata Example

## Work flow

1. Specify the survey design characteristics:

```
svyset su1 [pw=pw], strata(st1) ...
```

2. Estimate the population parameter of interest:

```
svy: mean y
```



# Stata Example

## Postratification

- ▶ Using `f`

```
svyset su1 [pw=pw], strata(st1)    ///  
      poststrata(f) postweight(w4f)
```

# Stata Example

## Raking-ratio using factor variable `f`

- ▶ Without population size, need `bn.`

```
svyset su1 [pw=pw], strata(st1)          ///  
      rake(bn.f, totals(1.f=45931      ///  
                2.f=54772           ///  
                3.f=52677           ///  
                4.f=46620))
```

- ▶ With population size, `i.` is sufficient

```
svyset su1 [pw=pw], strata(st1)          ///  
      rake(i.f, totals(1.f=45931      ///  
                2.f=54772           ///  
                3.f=52677           ///  
                4.f=46620           ///  
                _cons=200000))
```

# Stata Example

## zero non-response sample, using **f**

Variable	orig	post	rake	regress
y	53.005247	62.788326	62.788326	62.788326
	7.4721232	5.3039955	5.3039955	5.3039955
N_pop	200,000	200,000	200,000	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Weight adjustment changed the point estimate.
- ▶ Smaller variance estimates indicate a more efficient mean estimate.

# Stata Example

## zero non-response sample, using **g**

Variable	orig	post	rake	regress
y	53.005247	54.091047	54.091047	54.091047
	7.4721232	6.8654765	6.8654765	6.8654765
N_pop	200,000	200,000	200,000	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Recall that **g** is not as strongly associated with **y** as **f**.
  - ▶ Smaller change to the mean estimate.
  - ▶ Smaller change in the variance estimates.

# Stata Example

## Raking-ratio using factor variables **f** and **g**

```
svyset su1 [pw=pw], strata(st1)    ///  
    rake(bn.f bn.g,                ///  
          totals(1.f=45931         ///  
                 2.f=54772         ///  
                 3.f=52677         ///  
                 4.f=46620         ///  
                 1.g=98757         ///  
                 2.g=101243) )
```

# Stata Example

## zero non-response sample, using **f** and **g**

Variable	original	rake	regress
y	53.005247 7.4721232	64.435965 4.2315801	64.079348 4.2355881
N_pop	200,000	200,000	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Distinct mean estimates.
- ▶ Bigger reduction in the variance estimates.

# Stata Example

## Raking-ratio using continuous variable **a**

- ▶ Using **a** without population total

```
svyset su1 [pw=pw], strata(st1)      ///  
      rake(a, totals(a=2000000))
```

- ▶ Using **a** with population total

```
svyset su1 [pw=pw], strata(st1)      ///  
      rake(a, totals(a=2000000      ///  
                  _cons=200000))
```

# Stata Example

zero non-response sample, using **a**

Variable	orig	rake_noc	rake
y	53.005247 7.4721232	60.855469 3.6519173	64.083179 3.6369672
N_pop	200,000	218,098	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Distinct mean estimates.
- ▶ Big reduction in the variance estimates.
  - ▶ Recall the strong association between **y** and **a**.



# Stata Example

zero non-response sample, using **b**

Variable	orig	rake_noc	rake
y	53.005247	52.43749	52.399275
	7.4721232	6.4023137	6.4042111
N_pop	200,000	199,239	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Recall that **b** is not as strongly associated with **y** as **a**.

# Stata Example

## Calibration

- ▶ Using **a** and **b**

```
svyset su1 [pw=pw], strata(st1)      ///  
      rake(a b, totals(a=2000000    ///  
                  b=1000000        ///  
                  _cons=200000))
```

# Stata Example

zero non-response sample, using **a** and **b**

Variable	orig	rake	regress
y	53.005247	63.553724	63.613031
	7.4721232	1.5635263	1.5635551
N_pop	200,000	200,000	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Distinct mean estimates.
- ▶ Biggest reduction in the variance estimates.

# Stata Example

## Sample from the population, with non-response

Stratified two-stage design:

1. select 20 PSUs within each stratum
2. select 10 individuals within each sampled PSU

With 10% non-response, this sampling scheme yielded:

- ▶ 361 sampled individuals
- ▶ constant sampling weights

$$pw = 500$$

# Stata Example

## Sample weighted cell totals for f

```
. table f [pw=pw], c(freq min w4f) format(%9.0gc)
```

f	Freq.	min(w4f)
1	18,500	45,931
2	66,500	54,772
3	44,500	52,677
4	51,000	46,620

- ▶ Over-represented: 2, 4
- ▶ Under-represented: 1, 3

# Stata Example

10% non-response sample, using **f**

Variable	orig	post	rake	regress
y	68.335883 6.6819885	63.452068 5.5113469	63.452068 5.5113469	63.452068 5.5113469
N_pop	180,500	200,000	200,000	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Weight adjustment changed the point estimate.
- ▶ Smaller variance estimates, as we expected.

# Stata Example

## 10% non-response sample, using **f** and **g**

Variable	orig	rake	regress
y	68.335883	59.974513	60.234483
	6.6819885	4.4071179	4.464893
N_pop	180,500	200,000	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Distinct mean estimates.
- ▶ Bigger reduction in the variance estimates.

# Stata Example

10% non-response sample, using **a**

Variable	orig	rake	regress
y	68.335883 6.6819885	58.572179 4.3092797	58.595651 4.3223863
N_pop	180,500	200,000	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Distinct mean estimates.
- ▶ Big reduction in the variance estimates.
  - ▶ Recall the strong association between **y** and **a**.



# Stata Example

10% non-response sample, using **a** and **b**

Variable	orig	rake	regress
y	68.335883 6.6819885	59.887132 1.1631547	59.885356 1.1586559
N_pop	180,500	200,000	200,000

legend: b/se

- ▶ Reminder:  $\mu_y$  is 60
- ▶ Distinct mean estimates.
- ▶ Biggest reduction in the variance estimates.

# Summary

- ▶ Calibration weight adjustments are determined by the original sampling weights and auxiliary variables.
- ▶ Expect more efficient estimates for outcomes that have a strong association with the auxiliary variables.
- ▶ Use **svyset** option **rake()** or **regress()**.
  - ▶ Use **bn.** operator for factor variables in *varlist*.
  - ▶ Use **\_cons** to specify the population size in **totals()**.
- ▶ Use **svy:** prefix.
  - ▶ All variance estimation methods support calibration.

# References

Deville, J.-C., and C.-E. Särndal. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87: 376–382.

Deville, J.-C., C.-E. Särndal, and O. Sautory. 1993. Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* 88: 1013–1020.

# References

Lumley, R., P. A. Shaw, and J. Y. Dai. 2011. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review* 79(2): 200–220.

web: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3699889>

Valliant, R. 2002. Variance estimation for the general regression estimator. *Survey Methodology* 28: 103–114.

Valliant, R., and J. Dever. 2018. *Survey Weights: A Step-by-Step Guide to Calculation*. College Station, TX: Stata Press.